ANALYSING POLYNUCLEOTIDE SEQUENCES

1. INTRODUCTION

Three methods dominate molecular analysis of nucleic acid sequences: gel electrophoresis of restriction fragments, molecular hybridisation, and the rapid DNA sequencing methods. These three methods have a very wide range of applications in biology, both in basic studies, and in the applied areas of the subject such as medicine and agriculture. Some idea of the scale on which the methods are now used is given by the rate of accumulation of DNA sequences, which is now well over one million base pairs a year. However, powerful as they are, they have their limitations. The restriction fragment and hybridisation methods give a coarse analysis of an extensive region, but are rapid; sequence analysis gives the ultimate resolution, but it is slow, analysing only a short stretch at a time. There is a need for methods which are faster than the present methods, and in particular for methods which cover a large amount of sequence in each analysis.

This invention provides a new approach which produces both a fingerprint and a partial or complete sequence in a single analysis, and may be used directly with complex DNAs and populations of RNA without the need for cloning.

In one aspect the invention provides apparatus for analysing a polynucleotide sequence, comprising a support and attached to a surface thereof an array of the whole or a chosen part of a complete set of oligonucleotides of chosen lengths, the different oligonucleotides occupying separate cells of the array and being capable of taking part in hybridisation reactions. For studying differences between polynucleotide sequences, the invention provides in another aspect apparatus comprising a support and attached to a surface thereof an array of the whole or a chosen part of a complete set of oligonucleotides of chosen lengths comprising the polynucleotide sequences, the different

E SOUTHERN - USSN 08/230,012

APPLICANTS' COMMENTARY

Introduction

The application in suit relates to arrays of oligonucleotides and their use in investigating polynucleotide sequences. The application is based on a British application 8810400.5 filed 3 May 1988 and on an international application PCT/GB89/00460 filed 2 May1989. Professor E M Southern's invention has proved to be seminal. This introduction puts it in historical context.

Early applications of molecular hybridization to practical problems were greatly facilitated by the introduction by Gillespie and Spiegelman of a system in which one of the interacting nucleic adds was bound to a solid support. In the original method, DNA was bound to a nitrocellulose membrane, the other nucleic acid, DNA or RNA, was labelled and applied to the membrane bound DNA under hybridization conditions, and after washing away the unbound target, the extent of hybridization was measured from the radioactivity of the membrane, estimated by scintillation counting or Cherenkov counting. The method found very many applications and was extended in a number of ways. The most enduring adaptations have been the so called blotting methods, Southern, northern and dot-blotting which are still in widespread use for the detection of specific nucleic acid sequences.

In the mid 1980s, a number of scientists most notably Renato Dulbecco, Ayishi Wada and James Watson promoted the idea of sequencing the human genome. Several international meetings were held to discuss ways and means of achieving this ambitious goal. Wada organised one of these meetings in Okayama in July 1986. At that

meeting Southern heard a colleague of Wada's describe his work on the application of column-bound oligonucleotides to analyze mutations which showed that conditions could readily be found under which a perfectly matched duplex was stable and a single mismatch caused complete abolition of duplex formation. There is no difference in principle between this result and earlier work of Bruce Wallace, in which the DNA to be analysed was bound to the solid phase and the labelled oligonucleotide used as solution phase probe. However, the fact that the oligonucleotide was bound to the solid phase suggested to Southern that many different oligonucleotides could be tested against a single labelled target. Southern was struck by the potential of this simple method, realising that the test could be used for much more than simply distinguishing a single, characterised mutant from its wild type counterpart sequence. A positive interaction between an oligonucleotide of known sequence and a target effectively "read" that part of the sequence with which it formed duplex, through the Watson-Crick rules of base pairing, and he saw how the method could be adapted to solve several problems. The different analytical approaches that he envisaged included sequence determination, sequence comparison for the analysis of unknown mutations, sequence comparison for the analysis of mRNA populations, mutation screening, linkage analysis and genetic fingerprinting; these ideas are outlined in the specification of the application in suit.

Though inventive, this insight would have been fruitless if there were no practical way of implementing it, and the true originality of Southern's application can only be seen by considering the solutions he presents to the practical problem of providing oligonucleotides, well-defined chemically, in a format which permits several to be presented to a target simultaneously.

During the years following the subject patent application, very many different designs of oligonucleotide arrays have been proposed.

These can be envisaged as intermediate between two extremes, which will hereafter be called the sequencing mode and the diagnostic mode:-

The sequencing mode, in its pure form, involves the use of a complete set of oligonucleotides of a given length in the form of an array to investigate the sequence of a polynucleotide of wholly unknown sequence. The size of the array depends on the spacing between adjacent cells and their number which in turn depends on the chosen oligonucleotide length. Analysis of the results is generally computer-assisted.

The diagnostic mode, in its pure form, involves the use of two oligonucleotides, immobilised at spaced locations on a surface, to analyze a variant of a predetermined polynucleotide sequence. For example, two immobilised oligonucleotides which differ by a single nucleotide can be used to discriminate between two alleles of a gene.

Between the pure sequencing mode and the pure diagnostic mode, there are many intermediate modes in which this invention can be practised. These may involve analysis of sequence variation where the target sequence is partly undetermined.

The Sequencing Mode

In the field of DNA sequence analysis, The Human Genome Project has provided a need for more efficient methods; in particular, methods which are readily automated. Many new methods have been suggested but few have survived preliminary trial. However, Southern and his colleagues have shown the feasibility of sequencing by hybridization to arrays of oligonucleotides in a paper published in 1992 in *Genomics,* the main outlet for papers relating to the Human Genome Project (Genomics, 13, 1008-1017, 1992). In this paper, Southern and colleagues described a number of developments. They developed novel methods for synthesising oligonucleotides on glass (EPA 386 229) and plastic surfaces (Beckman patent) and novel combinatorial methods (described in the application in

suit) for creating arrays of several thousand oligonucleotides of chosen length and sequence within a relatively small area that could be hybridised with a labelled target sequence in a single simple operation that is readily automated. They devised ways of collecting the large amount of data produced from such an analysis. They developed sophisticated statistical methods for analysing the data to produce a sequence from the hybridization signals.

The Genomics paper remains as the best demonstration of the application to sequence determination. However, at least ten other research groups and several companies, recognising the importance of the method, have entered the field. These other groups have concentrated on devising ways of creating arrays of large number of oligonucleotides within a small area on a solid support. There have been three approaches based on ideas described in the application in suit. One approach, used by several of the groups, has been to adapt ink-jet printer technology to deliver oligonucleotide precursors to the surface in small droplets; so far this approach has not produced commercially useable arrays. Two other approaches use masking techniques to confine either the precursors or the deprotecting agents to defined areas during synthesis.

The most sophisticated of the masking methods required the development of novel nucleotide reagents with photocleavable protecting groups, so that the techniques of photolithography used in the semiconductor industry could be adapted to making arrays. Tens of millions of dollars have been invested in this development, which now seems on the brink of successful application (BioTechniques, 19, 442-7, 1995). Physical masking methods for confining the coupling agents, capable of synthesising several thousand oligonucleotides on a glass, silicon, or plastic surface, have been developed in Professor Southern's laboratory, by Beckman Instruments and by Affymetrix.

The most novel approach, developed in Professor Southern's

laboratory, uses electrically addressable arrays of microelectrodes to generate acid electrochemically (WO 93/22480). This is applied at the deprotection step to create patterns of oligonucleotides on the surface of a substrate offered to the electrode array. This procedure has been shown to be capable of generating cells as small as 2 microns, opening up the possibility of generating millions of oligonucleotides within an area of a few square millimeters.

The hybridization step is not difficult, but new methods are needed to more efficiently detect and measure the extent of hybridization and much time, effort and several millions of dollars of investment have gone into this stage of the process. Affymetrix have developed a system using confocal microscopy with fluorescent labelling of the target sequence. They are continuing this development in collaboration with Hewlett Packard and Molecular Dynamics. Roche have an array-based colorimetric DNA fingerprinting technique that was used in the O J Simpson trial. Abbott have developed a diagnostic method in which the target is labelled with Se particles which are detected by light scattering of an evanescent wave. Genometrix, in collaboration with the Lincoln Laboratories at MIT, have explored techniques for directly measuring the intrinsic electrical properties of the target DNA, so that the array would be created on an integrated semiconductor device. It has been claimed that “DNA chip” technology, according to the application in suit, will have sales in excess of one billion dollars by the year 2002.

There have been several research publications on the theory of sequencing by hybridization and ingenious ways have been developed for extracting sequence information from the hybridization data, including the method developed by John Elder in Southern's laboratory already referred to.

The topic of sequencing by hybridization has been the subject of reviews in the scientific literature; it is a regular topic at scientific

meetings on DNA sequencing, indeed there are annual meetings dedicated to the topic.

Several patents claiming improvements, extensions and implementations of the method have been filed.

Two US Government funding agencies have invested a total of almost $40M towards research in the technology. Thus, there is every indication of vigorous activity and commercial interest in sequencing by hybridization.

Analysis of Sequence Variation

There is a large and growing need for an improved method for analysing sequence variation, again fuelled by large scale genetic programmes such as the Human Genome Project

There are three stages in such a programme where sequence variation is analysed:

1. Linkage mapping using DNA markers requires the analysis of many thousands of markers against a panel of tens to hundreds of individuals in families or pedigrees. At present this is done by gel based methods which are very labour intensive. The procedure must be repeated for every genetic character that is mapped. Thus the scale and cost of this operation is vast.

2 Once linkage has determined the position of a gene of interest between two DNA markers, the next step is to isolate all the coding sequences in the region and to compare them in wild-type and mutant individuals to find the gene associated with the genetic trait. There may be hundreds of genes, each several kilobases in length, that have to be analysed in many individuals. Again, this is an enormous and labour intensive task. Several alternative methods are used, but almost all of them require gel electophoresis. The fact that an annual meeting is held to discuss technical progress in the development of new methods indicates

the level of dissatisfaction with them.

3. When the gene associated with a particular trait has been isolated and sequenced. All families carrying the gene are analysed to determine the spectrum of mutations. The number of different mutations is often large: there are around 500 different mutations in the CFTR gene associated with cystic fibrosis.

4. Once the spectrum of mutations is understood, standard tests can be developed to detect them. Such tests are applied to individuals at risk of carrying a disease gene, for example. They may also be used to genotype individuals in animal or plant breeding programmes. Many of these programmes must be carried out on a very large scale.

Southern and his colleagues have contributed several array-based methods for the analysis of sequence variation. The most general method, which can detect differences in sequences which are related by point mutations can be applied to sequences which are not predetermined and, therefore, this method which has been shown to work in model system, can be used in all of the applications described above; it is described in the *Genomics* paper referred above.

The Diagnostic Mode

This is in essence a reverse of the known dot-blot format and is applied to sequences and variants which are already known. In this diagnostic mode also, it is important for practical operation that a single mismatch have a substantial destabilising effect on the duplex. The extent of this destabilising effect depends on various factors, including the lengths of the oligonucleotide probes and the position of the mismatch in them. Applicants have developed a technique of synthesising oligonucleotide probes on a surface in the form of an array ready for use. This technique has advantages: the length of each oligonucleotide can be precisely controlled and varied; the position of a prospective mismatch can be

controlled; these variables can readily be adjusted to determine the optimum conditions for efficient operation.

The importance of this approach is clearly illustrated in Example 3 and in Table 1 of the subject application. These are arrays intended, not for practical diagnosis, but to study the effect of mismatches, by varying the length of the oligonucleotide and the position of the mismatch in the sequence, in order to permit the design of arrays optimised for use in a diagnostic mode. Longer oligonucleotide probes do not necessarily give better discrimination than shorter probes. For any given length of probe, the position of a single nucleotide mismatch is important in determining discrimination at a given temperature.

This approach results in oligonucleotide arrays having various features:-

a) Each oligonucleotide is attached to the surface of the support through a terminal residue. This results automatically from applicants' preferred technique of synthesising oligonucleotides *in situ* on a surface of the support. It also ensures that all the bases of the oligonucleotide molecules are capable of taking part in hybridization reactions, a result which is not ensured if the attachment method does not guarantee terminal linkages.

b) Each oligonucleotide is attached to the surface of the support through a covalent linkage. This results from applicants' preferred technique of synthesising the oligonucleotides *in situ* on a surface of a support. Covalent attachment ensures that oligonucleotides are not lost during hybridization; such losses would make the analysis unreliable.

c) Each oligonucleotide has a chosen length and sequence which can readily be optimised to discriminate between matches and mismatches and which may be in the range of 8 to 20 nucleotides, although longer oligonucleotides may also be useful in the diagnostic mode.

There are many ways of applying the diagnostic mode to practical problems. One method uses arrays which are dedicated to a particular target sequence; arrays of this type are described in the application in suit. Southern and his colleagues have devised a way of making such arrays; the method produces sets of overlapping oligonucleotides which represent the complements of all positions of the target sequence. A particularly attractive feature of the method is that oligonucleotides of all lengths up to a chosen maximum are made on the array. Such arrays, which can be used to scan the hybridization behaviour of the full length of the target sequence and thus detect mutation at any position, are described in Nucleic Acids Research, 22, 1368-73, 1994. Affymetrix have developed similar arrays, but their design incorporates all possible single base variants of the target sequence in the oligonucleotides on the array; an array designed to analyze mutation in HIV is described in BioTechniques, 19, 442-7, 1995.

Another approach (developed in Southern's laboratory) allows the simultaneous analysis of a large number of individuals with a similar number of allele specific oligonucleotides (ASOs). It involves creating arrays of stripes of oligonucleotides which can then be addressed by applying target sequence in orthogonal stripes (Nucleic Acids Research, 21, 2269-70, 1993). The great benefit of this combinatorial approach is that the number of analyses is equal to the product of the number of ASOs and the number of targets applied. Thus with a hundred ASOs and an equal number of targets, a total of 10 000 analyses could be carried out simultaneously, saving a vast amount of work.

Amended Claims

Arising out of this, applicants wish to amend their claims. The claims on file are deleted and are replaced by the attached set, in

which independent claims are generally directed either to: analysing undetermined sequences (the sequencing mode) or undetermined sequence variants of polynucleotides; or to analysing predetermined polynucleotide sequences (the diagnostic mode).

Applicants are filing separately an Information Disclosure Statement in which the following reference is highlighted.

D1, EPA 0 235 726, Datta Gupta

This reference D1 describes the use of immobilised oligonucleotides as probes in what is essentially a reverse dot-blot format. The disclosure concerns the diagnostic mode of analysis, i.e. the analysis of known sequence differences in known sequences; it is not relevant to any claim of the subject application concerned with the sequencing mode.

One object of the invention is to provide a method for distinguishing alleles of individual genes (page 3 line 25). According to the invention (page 3 lines 50 to 54):-

"A labelled nucleic acid test sample is contacted simultaneously with several different types of DNA probes for hybridization. The nucleic acid test sample is labelled and hybridised with several unlabelled immobilised probes. The positions of the probes are fixed, and the labelled probe detected after hybridization will indicate that the test sample carries a nucleic acid sequence complementary to the corresponding probe."

Page 4 line 35 to page 9 line 5 describe methods of labelling nucleic acids. Page 9 line 6 to page 10 line 31 describe immobilised nucleic acid probes and various assay formats involving them.

The disclosure at page 9 lines 25 to 40 contains a discussion about covalent or non-covalent bonding of oligonucleotides to supports. As described at lines 30 to 33, covalent binding involves the use of intermediate compounds that react with the bases of oligonucleotide

residues; so any covalent binding would not necessarily be by a terminal nucleotide residue. The paragraph at lines 34 to 40 describes non-covalent immobilisation achieved by phosphorylation, but it is not clear whether or how this would work.

The following passage appears at page 10 lines 32 to 40:-

"The present invention relates to a novel hybridization technique where probes are immobilised and an eukaryotic nucleic acid sample is labelled and hybridised with immobilised unlabelled probe. A surprising characteristic of the invention is the ability to detect simple or multiple copy gene defects by labelling the test sample. Since there is no requirement for an excess of labelled hybridising sequence, the present method is more specific. In the present invention, simultaneous detection of different gene defects can be easily carried out by immobilising specific probes.

For example, using the present invention, one can immobilise oligonucleotide probes specific for genetic defects related to sickle cell anemia and probes for alpha-thalassemias on a sheet of nitrocellulose paper, label the test sample and hybridise the labelled test sample with the immobilised probes."

The introductory part of the specification, from page 2 to page 13 line 9, is somewhat confusing because it contains statements inconsistent with the passages quoted above. Thus page 3 lines 28 to 31 appear to suggest that the probes are labelled and the unknown is unlabelled and immobilised. This impression is endorsed by the otherwise obscure sentence at page 3 line 47. The paragraphs at page 9 line 41 to 52 describe standard dot-blot and sandwich hybridization assays, yet are presented as being part of the invention. The sentences at page 12 lines 42 to 44 indicate that immobilised probes can be labelled.

This confusion in the introductory part of the specification carries over into the experimental part. Examples 6 and 7 are of interest

and are discussed in detail. Whatever the merits of D1 as a whole, these two examples are worthless. To see why, it is helpful to look at the sequence set out at page 19 line 57 of D1 and which is reproduced below:-

|<——16-mer ——>| T |<——16 mer ——>|

5' CTGCT$_4$AATCTTAAGGAT$_3$AAT$_4$CTCCTGAGGAGAAGTCTGCT$_4$AATCTTAAGGAT$_3$AAT$_4$CT 3'

|<————19-mer————>|

43-mer——>|<———— 43-mer ————>|<–43-mer

The 19-mer is (complementary to) the haemoglobin sequence under investigation. Where the normal haemoglobin allele has an A the sickle cell mutation has a T at the position shown. The 43-mer encompasses the 19-mer with additional chain at both ends. The 16-mer acts simply as a glue to hold molecules of the 43-mer in place for ligation.

In Example 6, three different techniques are used to immobilise oligonucleotides on a nitrocellulose or nylon support.

In method 1, 43-mers A and S (A = normal globin gene; S = sickle globin gene) were kinased using 32P-ATP; in other words, a 32P radioactive label was attached to the end of the 43-mer. Two dilutions of these labelled oligonucleotides were spotted onto nitrocellulose (NC) and nylon (NY) membranes.

In method 2, radiolabelled 43-mers were polymerised by means of a ligase before application to a membrane. The 16-mer was used to temporarily hold two 43-mers in place so that ligation can proceed. The products thus had lengths of multiples of 43 nucleotides; and were radiolabelled.

In method 3, the radiolabelled 43-mers were cross-linked prior to application to membranes. Thus the products were not single stranded oligonucleotides; but they were radiolabelled.

These radiolabelled products were spotted onto nitrocellulose or nylon membranes and baked at 80°C. The immobilised probes were

then hybridised with unknowns which are, in Example 6, not described in any detail. The results are shown in Figure 1 of D1. This is an autoradiograph which shows a number of spots at spaced locations. But it is not surprising that the spots are present, for all the probes immobilised were radiolabelled! The experiment and results give absolutely no information about the target oligonucleotide under investigation.

As an anticipation of applicants' claimed invention, Example 6 and Figure 1 of D1 fail because they employ immobilised oligonucleotide probes which are radiolabelled and therefore inherently incapable of generating any sequence information.

Example 7 describes a similar experiment, but one in which the unknown target in solution is photolabelled with BPA for non-radioactive detection. The results are shown in Figure 2 which is a photograph comprising four columns, each of nine rows which presumably correspond to the nine immobilised probes listed at page 19 lines 41 to 54. The unknown DNA under investigation should hybridise to the probes immobilised in rows 1, 2, 5, 7 and possibly 6, but not to the probes in rows 3, 4 and 9. The probes in row 5 are in effect a positive control to which everything should hybridise; row 8 gives a positive signal because the DNA that was spotted was labelled with biotin; the probes in row 9 are in effect a negative control to which nothing should hybridise. Stringency washes are carried out at four different temperatures as indicated by the four columns in Figure 2. The labelled unknown should remain immobilised at relatively low temperatures, but should be removed and disappear at relatively high temperatures.

Inspection of Figure 2 of D1 shows that something is seriously wrong. Hybridization has taken place in row 9, which it should not have. Hybrids which evidently have melted at 57° are stable at 60°; this should never happen. In short, Figure 2, and indeed the whole example, appears to have little meaning.

The technique described in Example 6 involved spotting radiolabelled probes on to nitrocellulose or nylon membranes which were then baked at 80°C. That is expected to produce an attachment of the products to the membrane which was both random and unreliable - as shown by the fact that some of the spots in Figure 2 are faint or missing. The products were definitely not attached to the membranes reliably and uniformly through a terminal nucleotide. The products were definitely not attached to the membranes reliably and uniformly through a covalent link. The products attached to the membranes were not of length chosen to discriminate between matches and mismatches, e.g. 8 to 20 nucleotides in length. Thus each of apparatus claims 41 to 43, of the subject application is novel over the disclosure.

The techniques described in Example 6 of D1 involve random attachment of oligonucleotides to supports. The result is inevitably that some molecules are capable of hybridising to complementary oligonucleotides, while others are stereochemically prevented from doing so. Thus the hybridising efficiency is inevitably low. One may speculate that the authors of D1 used 43mers in an attempt to overcome this problem - the longer the oligomers, the greater is the chance that the relevant 19 nucleotide section thereof may be stereochemically free to hybridise with its complementary molecule. Also, preferred 19-mers could not readily be immobilised on a membrane, even when phosphorylated.

As noted, the authors of D1 needed to use 43-mers or longer oligonucleotides in order to get them to immobilise on the support while still retaining reasonable hybridising efficiency. But 43-mers are almost certainly not the optimum length to achieve maximum Tm (melting temperature) discrimination between exact matches and single mismatches. So the authors of D1 put a specific 19mer for detecting the sickle cell anaemia mutation in their 43mer. But the length of the 19mer may not be optimum for the purpose; and the rest of the 43mer is at best

irrelevant to the hybridization reaction, but could interfere by producing a false signal, if, by chance, it contained sequences that were complementary to others in the target sequence, as could well be the case in complex sequences such as the whole genomic DNA analysed in the example. Indeed, applicants found that 15mers were optimum, and certainly better than 19mers for this particular sequence (Maskos & Southern, Nucleic Acids Research, 1993, Vol 21, No 9, 2267-8). By contrast, applicants' preferred technique, involving the in situ synthesis of oligonucleotides of chosen length bound by a covalent link through a terminal nucleotide to a surface of a support, permits control of all these factors and makes it easy to determine optimum conditions and oligonucleotide lengths and sequences. These are real technical advances which characterise applicants' invention over D1.